

DP18 – Feedback Loops and Reputation

The Conditions for Learning, Recognition, and Repair

1. Purpose of This Draft

This draft articulates Desirable Property 18 (DP18) as the condition under which the meta-layer can learn from participation without collapsing into surveillance, popularity contests, social credit, or reputation systems that permanently trap people in old contexts.

DP18 defines how communities gather feedback, convert signals into accountable learning, and recognize trustworthy contribution over time. It treats feedback and reputation as civic infrastructure, not engagement metrics.

Feedback loops are how the meta-layer senses whether its governance, incentives, safety systems, interfaces, and community norms are working. Reputation is how the system remembers patterns of contribution, care, reliability, and harm without reducing people to a single score.

If DP18 is weak, predictable failures follow: communities ask for feedback but do not act on it; reputation becomes a vanity metric; harmful actors launder trust across contexts; contributors burn out because recognition is invisible; AI systems optimize for shallow approval signals; and governance adapts only after legitimacy breaks.

DP18 connects directly to:

- DP1, identity and accountability
- DP2, participant agency and empowerment
- DP3, adaptive governance
- DP4, data sovereignty and privacy
- DP7, interoperability
- DP8, community-defined participation and governance zones
- DP9, developer and community incentives
- DP12 and DP13, AI governance and containment
- DP14, transparency and trust
- DP20, ownership and stewardship

DP18 does not prescribe a universal reputation score, a single feedback interface, or one model of rewards. It defines the minimum conditions under which feedback and reputation remain contextual, contestable, privacy-preserving, and aligned with community flourishing.

2. Problem Statement

Today's web is saturated with signals, but starved of trustworthy feedback.

Likes, shares, followers, ratings, badges, view counts, and engagement graphs appear to measure social value. In practice, they often measure visibility, habit formation, emotional provocation, automation, or the ability to game a platform's ranking system.

This produces recurring failures:

- communities provide feedback but cannot see whether it changed anything
- bad actors accumulate credibility through volume, performance, or network effects
- positive contributions disappear into feeds without durable recognition
- reputation is trapped inside platforms and lost when people move
- marginalized participants are excluded when feedback systems privilege dominant voices
- AI systems learn from shallow approval signals rather than community-defined values
- governance becomes reactive because weak signals are ignored until crisis

DP18 reframes feedback and reputation as learning infrastructure. Feedback must produce visible adaptation. Reputation must be contextual memory, not global judgment.

A healthy meta-layer must be able to answer:

- What did participants signal?
- Who or what is being evaluated?
- What changed as a result?
- Which signals are reliable in this zone?
- Which signals are being gamed?
- How can a participant contest, repair, or outgrow a reputation state?
- How does the system prevent feedback from becoming surveillance or social punishment?

Without DP18, the meta-layer cannot mature. It can collect comments, votes, and metrics, but it cannot reliably learn.

3. Threats and Failure Modes

3.1 Feedback theater

Communities are asked for input, but feedback has no visible path into decisions or operations.

Example: A town hall gathers participant concerns, but the same interface defaults, moderation rules, and incentive structures remain unchanged without explanation.

Why this matters: Feedback without response erodes trust faster than silence because it simulates agency while preserving control.

3.2 Reputation as popularity

Reputation collapses into follower counts, visibility, or applause.

Example: A participant becomes trusted because their posts receive high engagement, even though their contributions are often misleading or inflammatory.

Why this matters: Popularity is not reliability. DP18 requires multidimensional, context-bound reputation.

3.3 Runaway feedback loops

A signal increases visibility, visibility increases signal volume, and the system treats the resulting amplification as legitimacy.

Example: Early upvotes push a submission into prominence; prominence generates more upvotes; the system mistakes attention momentum for quality.

Why this matters: Feedback loops must be bounded. Otherwise reputation becomes a mechanism for self-reinforcing inequality or manipulation.

3.4 Reputation laundering

Actors transfer trust from one context into another where it has not been earned.

Example: A participant with high reputation in a gaming community receives influence in a medical advice zone without relevant credentials or history.

Why this matters: Reputation must preserve context, scope, and provenance. Portability without boundaries becomes a trust exploit.

3.5 Permanent stigma and no repair path

Negative reputation becomes permanent, opaque, or disproportionate.

Example: A participant makes a mistake in one zone and is silently down-ranked everywhere without notice, appeal, or decay.

Why this matters: Accountability must support learning and repair. Systems that never forgive incentivize abandonment, evasion, and identity cycling.

3.6 Feedback capture by loud minorities

Coordinated or highly resourced groups dominate feedback channels.

Example: A small faction floods surveys, flags, votes, or comments to steer governance toward its preferred outcome.

Why this matters: Feedback systems must distinguish broad legitimacy from coordinated pressure.

3.7 Marginalized voice suppression

Feedback mechanisms reproduce structural exclusion.

Example: Town halls occur in one language and time zone, while reputation rewards favor participants already fluent in dominant cultural norms.

Why this matters: Adaptive governance requires feedback from those most affected, not only those most available or socially rewarded.

3.8 AI-optimized approval hacking

AI systems learn to maximize visible approval signals rather than community-defined value.

Example: An assistant generates emotionally pleasing but low-integrity responses because users reward confidence and fluency more than accuracy.

Why this matters: AI feedback loops must be evaluated against zone policy, truthfulness, safety, and long-term outcomes, not only immediate satisfaction.

3.9 Privacy erosion through reputation telemetry

Reputation systems collect excessive behavioral data to infer trustworthiness.

Example: A system tracks every interaction, dwell time, private message, and social association to generate reputation scores.

Why this matters: Trust cannot be built by violating data sovereignty. DP18 must operate under DP4 constraints.

3.10 Role ossification

Reputation grants access or authority, then authority produces more reputation, making roles difficult to contest.

Example: Early stewards accumulate status and retain control even as their contribution quality declines.

Why this matters: Reputation-linked roles require decay, review, rotation, and contestability.

3.11 Cross-system signal degradation

Feedback and reputation signals move across systems but lose meaning.

Example: A five-star rating exports without information about rubric, rater identity class, zone norms, time horizon, or dispute status.

Why this matters: Interoperable reputation must preserve semantic context or visibly degrade.

4. Core Principle

Feedback loops and reputation in the meta-layer must enable communities to learn from lived experience, recognize trustworthy contribution, correct harmful behavior, and adapt governance over time while preserving agency, privacy, context, contestability, and the possibility of repair.

Reputation is not a universal score. It is contextual memory under governance.

Feedback is not engagement. It is a structured pathway from signal to learning to action.

Example: A community receives recurring feedback that a moderation rule is being applied unevenly. The system clusters the feedback, surfaces affected groups, opens a review process, publishes a decision receipt, updates the rule, and tracks whether the change improves outcomes over time.

What this feels like: Participants can see that their experience matters, not because every request is granted, but because every meaningful signal has a legitimate path into community learning.

Without this: The meta-layer repeats the failure of today's platforms: it extracts signals from people while denying them visible influence over the system those signals shape.

5. Primary Mechanisms and Structural Conditions

5.0 Feedback and Reputation Layer: Signal, Memory, Adaptation, and Recourse

DP18 requires a feedback and reputation layer that converts participant signals into governed learning.

This layer includes:

- feedback objects
- reputation objects
- signal provenance
- confidence and uncertainty metadata
- decay and repair mechanisms
- dispute and appeal pathways
- adaptation receipts
- AI-assisted pattern detection under human ratification

The layer must operate at the interface where participants act and are affected, while preserving the governance and privacy boundaries of each zone.

A feedback and reputation layer is not a database of social judgments. It is an accountable learning system.

Failure mode: **social credit drift**, where contextual feedback becomes generalized behavioral scoring.

5.1 Feedback objects

Feedback must be represented as structured objects, not only comments or reactions.

A feedback object SHOULD include:

- subject: what is being evaluated
- signal type: report, endorsement, correction, rating, concern, gratitude, appeal, observation
- scope: zone, artifact, participant, agent, policy, interface, or process
- evidence link, when appropriate
- privacy level
- urgency and severity
- confidence level
- submitter role or credential class, where relevant
- timestamp and version context
- requested action, if any

This allows feedback to be routed, aggregated, audited, and acted upon without flattening all signals into likes or complaints.

Failure mode: **signal mush**, where all feedback becomes undifferentiated noise.

5.2 Feedback routing and response commitments

Feedback systems must define what happens after feedback is submitted.

For each class of feedback, systems SHOULD specify:

- who or what receives it
- expected response time
- escalation path
- visibility rules
- criteria for closure
- appeal or reopening pathway

Participants do not need to control every outcome, but they must be able to see whether feedback entered a legitimate process.

Failure mode: **black-hole feedback**, where reports, suggestions, and concerns vanish without trace.

5.3 Continuous feedback mechanisms

DP18 requires ongoing feedback channels, not one-time consultation.

Mechanisms MAY include:

- surveys
- town halls
- asynchronous deliberation
- embedded interface prompts
- post-decision retrospectives
- safety reports
- participatory audits
- contributor reviews
- community health dashboards
- affected-group listening sessions

Continuous feedback must be designed for accessibility across language, time zone, ability, bandwidth, and technical literacy.

Failure mode: **episodic listening**, where communities are consulted only during crisis or launch moments.

5.4 Adaptive feedback systems

AI and automation may support feedback systems by summarizing large volumes of input, detecting patterns, identifying anomalies, and surfacing underrepresented concerns.

However, adaptive systems MUST remain bounded by:

- visible purpose
- zone-defined policy
- privacy constraints
- human review for material decisions
- auditability of summarization and classification
- clear disclosure of AI involvement

AI may help communities see patterns. It must not quietly decide whose experience counts.

Failure mode: **automation capture**, where AI summaries become de facto governance decisions.

5.5 Reputation as contextual memory

Reputation must be scoped to context, contribution type, and governance zone.

A reputation object SHOULD include:

- subject identity or agent reference
- reputation domain, such as reliability, care, expertise, stewardship, safety, responsiveness, accuracy, generosity, or follow-through
- zone scope
- source signals
- weighting logic
- confidence level
- decay schedule
- dispute status
- portability constraints
- rights or capabilities affected

This prevents reputation from becoming a single global score while still allowing communities to remember meaningful patterns.

Failure mode: **global reputation collapse**, where one number pretends to summarize a person, agent, or organization across all contexts.

5.6 Reputation-based compensation and recognition

Reputation may inform compensation, rewards, access, visibility, and recognition, but only under explicit governance.

Reputation-linked compensation SHOULD satisfy the following conditions:

- metrics and rubrics are published
- rewardable contributions are defined
- anti-gaming rules are active
- human review exists for high-value outcomes
- rewards do not depend solely on popularity
- appeals are available
- contributors can understand how reputation affected compensation

Positive contribution should be rewarded without creating a system where people perform for metrics rather than serve the community.

Failure mode: **reputation farming**, where actors optimize for visible reward signals rather than actual value.

5.7 Dynamic role-based access

Roles and capabilities may adjust based on reputation, contribution history, and trust signals.

Examples include:

- moderation privileges
- proposal rights
- voting eligibility
- grant review access
- amplification capacity
- trusted annotator status
- steward responsibilities
- AI-agent deployment permissions

Role changes MUST be:

- explainable
- scoped to a zone
- revocable
- reviewable
- subject to decay or renewal
- protected against sudden manipulation

Reputation may open doors, but it must not create unaccountable hierarchy.

Failure mode: **role capture**, where reputation-linked access becomes permanent power.

5.8 Reputation decay, renewal, and repair

Reputation must change over time.

Systems SHOULD support:

- positive reinforcement for sustained contribution
- decay for inactivity or stale signals
- faster decay for low-confidence signals
- repair pathways after mistakes or sanctions
- restorative processes where communities choose them
- expiry of context-specific penalties
- continued visibility of serious unresolved harms where appropriate

The goal is neither amnesia nor permanent stigma. The goal is governed memory.

Failure mode: **frozen reputation**, where old signals dominate present reality.

5.9 Bad behavior reporting and community moderation

Participants and communities must be able to report harm, abuse, manipulation, and bad behavior through structured channels.

Reporting systems SHOULD support:

- evidence submission
- safety-sensitive privacy
- protection from retaliation
- triage by severity
- pattern detection across incidents
- community-defined moderation workflows
- appeal and correction mechanisms
- transparency summaries where safe

Reports are feedback objects with safety consequences. They must be treated with care, not merged into generic sentiment metrics.

Failure mode: **weaponized reporting**, where reporting tools become harassment or governance capture instruments.

5.10 Positive contribution signaling

DP18 requires mechanisms to recognize good behavior, not only punish bad behavior.

Positive signals MAY include:

- gratitude
- endorsement
- attestation
- contribution receipts
- peer review
- successful mediation
- helpful annotation
- quality bridge creation
- accurate correction
- inclusive facilitation
- maintenance work
- long-horizon stewardship

Systems **SHOULD** make quiet, prosocial labor visible without forcing all care work into performance.

Failure mode: **negative-only memory**, where systems remember harm but fail to recognize repair, support, and stewardship.

5.11 Multi-dimensional reputation

Reputation must be plural.

A participant may be highly trusted for one function and untrusted for another. A contributor may be excellent at technical review but poor at facilitation. An AI agent may be strong at summarization but not authorized for dispute mediation.

Reputation dimensions **MAY** include:

- accuracy
- reliability
- safety
- expertise
- care
- responsiveness
- collaboration
- originality
- stewardship
- governance judgment
- bridge quality
- conflict resolution
- maintenance reliability

Failure mode: **single-axis trust**, where one form of contribution grants unrelated authority.

5.12 Interoperable reputation with bounded portability

Reputation should be portable where useful, but only with context preserved.

Portable reputation **MUST** carry:

- issuing zone
- signal type
- evaluation rubric
- timestamp
- decay or expiry

- confidence level
- dispute status
- portability permissions
- limitations on interpretation

When context cannot be preserved, systems **MUST** signal degradation.

Failure mode: **semantic stripping**, where reputation travels as a badge or score without the meaning needed to interpret it.

5.13 Community health dashboards

Communities **SHOULD** have dashboards that surface aggregate feedback and reputation patterns without exposing unnecessary personal data.

Dashboards **MAY** include:

- unresolved feedback volume
- response times
- repeated pain points
- moderation appeals
- participation diversity
- contributor recognition gaps
- role concentration
- reputation distribution
- unresolved harms
- experiment outcomes
- rollback triggers

Dashboards must be designed to support learning, not surveillance or public shaming.

Failure mode: **dashboard theater**, where metrics are displayed but do not affect decisions.

5.14 Rollback and experiment learning

Feedback loops should support safe experimentation.

When communities test policies, interfaces, incentives, or reputation mechanisms, experiments **SHOULD** define:

- hypothesis
- success indicators
- harm indicators

- review date
- rollback conditions
- affected participants
- decision authority
- public learning artifact

This allows communities to try new mechanisms without making every experiment permanent. Failure mode: **irreversible experimentation**, where communities bear the cost of failed changes without recourse.

5.15 Feedback receipts and adaptation receipts

Feedback submissions and system changes should generate receipts.

A feedback receipt MAY include:

- what was submitted
- when it was received
- where it was routed
- privacy and visibility status
- next process step

An adaptation receipt MAY include:

- what changed
- which feedback or evidence informed the change
- who approved it
- expected outcome
- review date
- appeal or contestation path

Receipts close the loop between voice and action.

Failure mode: **untraceable adaptation**, where systems change without participants knowing why.

6. Feedback, Reputation, and AI Systems

AI systems may participate in DP18 in multiple roles:

- as subjects of feedback
- as assistants in feedback processing

- as agents with their own scoped reputations
- as participants in governed zones where permitted
- as tools for detecting manipulation, abuse, and emergent patterns

DP18 requires that AI feedback loops be especially constrained because automated systems can scale signals, interpret signals, and act on signals faster than communities can deliberate.

6.1 AI as subject of reputation

AI agents, models, tools, and automated services **SHOULD** have scoped reputation profiles when they act in ways that affect participants.

These profiles **MAY** track:

- accuracy
- refusal appropriateness
- safety incidents
- responsiveness to correction
- policy compliance
- hallucination rates
- data-handling reliability
- emotional safety
- task performance
- escalation behavior

AI reputation must be tied to version, deployment context, and operator accountability.

Failure mode: **agent amnesia**, where an AI system causes repeated harm but each deployment is treated as isolated.

6.2 AI-assisted feedback synthesis

AI may summarize, cluster, translate, and analyze feedback, but must preserve dissent and uncertainty.

Systems **SHOULD** prevent AI from flattening minority concerns into majority sentiment.

A good synthesis identifies:

- recurring themes
- contested interpretations
- affected groups
- evidence gaps
- urgency
- possible actions

- confidence limits

Failure mode: **consensus hallucination**, where AI creates the appearance of agreement where disagreement remains.

6.3 Feedback loops for AI containment

Feedback from participants can help communities refine containment policies for AI agents.

Examples include:

- reports of manipulative behavior
- consent boundary violations
- failures to disclose automation
- unsafe recommendations
- emotional overreach
- unauthorized cross-zone action
- model behavior drift

Containment feedback should trigger policy review, audit, rollback, rate limits, or suspension where appropriate.

Failure mode: **containment lag**, where AI behavior changes faster than feedback systems can respond.

7. Privacy and Data Sovereignty Requirements

Feedback and reputation systems must comply with data sovereignty.

DP18 systems SHOULD practice:

- data minimization
- purpose limitation
- consent-bound collection
- separation of public recognition and private telemetry
- pseudonymous feedback where safe
- protection for vulnerable reporters
- deletion or expiry where appropriate
- access controls for sensitive reputation evidence
- clear notice when signals affect rights or compensation

Reputation must not become a pretext for continuous behavioral surveillance.

Failure mode: **trust through surveillance**, where the system claims safety by collecting more data than communities can legitimately govern.

8. Governance, Accountability, and Agency Surfaces

Feedback and reputation systems are governance systems. They must themselves be governable.

Communities SHOULD define:

- what counts as valid feedback
- what reputation dimensions matter
- who can issue attestations
- how signals are weighted
- how disputes are handled
- how repair works
- how decay works
- when reputation affects access or rewards
- when human review is required
- how algorithms are audited
- how changes are approved

This governance must be visible, versioned, and contestable.

Failure mode: **hidden reputation law**, where invisible formulas determine social standing and access.

8.1 Accountability surfaces

Governance claims about feedback and reputation must be checkable by the people they affect.

DP18 therefore expects named surfaces rather than general assurances:

- a published feedback charter per zone, stating what is collected, who reads it, and what response commitments apply
- a reputation register listing active reputation domains, their issuers, their rubrics, and the decisions they gate
- a decision log linking governance changes to the feedback, evidence, or incidents that prompted them
- a steward roster identifying who reviews disputes, on what timeline, and under whose oversight
- an audit trail for changes to weighting, thresholds, and automated triage
- an escalation route when a steward is themselves the subject of a complaint

Where a system cannot name the surface, it cannot claim the accountability.

8.2 Participant agency surfaces

Accountability without leverage produces informed powerlessness. DP18 requires that participants hold usable controls over their own signals and standing:

- see which reputation domains describe them, in which zones, and with what evidence
- correct, annotate, or contest an attestation, and see the outcome of that contest
- withhold or scope consent for reputation telemetry that is not necessary for a claimed function
- export contribution history and reputation artifacts under DP4 and DP7 terms
- request attenuation or repair after a defined period, and understand the criteria
- decline reputation-gated roles without losing baseline participation rights

Example: A contributor opens their reputation panel, sees a "reliability" attestation issued by a moderation steward, disagrees with the underlying evidence, files a challenge, and receives a dated decision with reasoning and an appeal option.

What this feels like: Standing is something you can inspect and argue with, not something applied to you.

Without this: Reputation becomes an administrative fact about people that they can neither read nor answer.

Failure mode: **consultation without leverage**, where participants are told how the system works but cannot act on anything they learn.

9. Incentives and Power Analysis

Feedback and reputation systems redistribute power whether or not they intend to. Whoever defines valid signals, sets weights, issues attestations, and controls decay controls who becomes visible, trusted, paid, and promoted.

DP18 therefore treats incentive design as a governance question rather than a tuning question.

9.1 Who gains power from feedback systems

- **Signal definers** decide what counts as evidence. If only structured tickets count, participants who communicate in conversation, voice, or a second language lose standing (DP21, DP23).
- **Attestation issuers** convert observation into durable claims. Concentrated issuance recreates gatekeeping under a new name.
- **Weight setters** determine whose experience moves a decision. Weighting is policy, not math.

- **Interface owners** decide what is easy to report and what requires effort. Friction is an allocation of voice.
- **Aggregators** who hold cross-zone reputation data gain leverage over zones that cannot verify their inferences.

9.2 Predictable incentive distortions

- **Metric substitution:** systems optimize for measurable approval instead of durable contribution, because approval is cheaper to collect.
- **Volume advantage:** participants with time and fluency accumulate reputation faster than participants with equal skill and less availability.
- **Recognition asymmetry:** visible work such as posting accrues reputation while invisible maintenance such as triage, mediation, and care does not.
- **Complaint deterrence:** where reporting is costly or exposing, harm goes unrecorded and the system reads silence as health.
- **Steward overload:** unfunded review capacity turns response commitments into backlog, and backlog into cynicism (DP17).
- **Score arbitrage:** where reputation unlocks compensation, actors invest in signal production rather than in the work the signal was meant to represent.

9.3 Counter-incentives DP18 expects

- fund review, mediation, and repair capacity as core functions rather than volunteer surplus
- credit maintenance and care work explicitly as reputation domains (DP9, DP20)
- weight signals from affected groups rather than from the loudest cohort
- cap the decision weight any single issuer can hold, and rotate issuance
- decay reputation so that standing reflects current participation rather than seniority
- prohibit reputation-linked rewards from depending on signals participants cannot inspect
- publish who benefits from each reputation domain, so that beneficiaries are legible

Example: A zone notices that its "helpfulness" domain is dominated by participants in one time zone. It rebalances by crediting asynchronous mentoring and by rotating the stewards who issue attestations.

Failure mode: **reputation as rent**, where early or well-positioned participants convert accumulated standing into permanent advantage that newer contributors cannot reach.

10. Community Signals Informing DP18

DP18 did not originate as an abstraction. It reflects recurring signals from communities that have lived through extractive feedback and punitive reputation.

Signals repeatedly expressed by participants, moderators, and maintainers include:

- "We answer surveys and nothing changes." Feedback is collected as sentiment, not as a governance input.
- "I don't know why my standing dropped." Reputation effects arrive without explanation or appeal.
- "One mistake follows me forever." Communities want accountability with a path back, not permanent exile.
- "The loudest people set the agenda." Feedback volume is mistaken for feedback representativeness.
- "Moderation work is invisible." Care, triage, and de-escalation carry weight in practice but not in recognition.
- "Reporting is dangerous." Participants withhold reports because exposure risk exceeds expected remedy.
- "My reputation is trapped." Contribution history cannot move between tools, so leaving means starting over (DP7).
- "Scores feel like credit ratings." Participants distinguish sharply between contextual trust and generalized scoring.
- "AI summarized us wrong." Communities want to see, and contest, how synthesis was produced (DP12, DP14).
- "We need to know if it worked." Participants want evidence that changes made in their name improved outcomes.

These signals converge on a single structural request: make the loop complete and make the memory fair. DP18 encodes that request as design conditions rather than as goodwill.

Community signals are inputs to DP18, not ratification of it. They are recorded here so that later revisions can be checked against the concerns that motivated the property.

11. Evaluation Criteria

A DP18-aligned implementation should be evaluated against the following questions.

11.1 Signal quality

- Are feedback signals typed, scoped, and attributable at the right level?
- Can systems distinguish evidence, opinion, endorsement, complaint, and appeal?

- Are low-confidence signals prevented from causing high-impact outcomes without review?

11.2 Loop closure

- Can participants see whether feedback was received, routed, reviewed, and acted upon?
- Do system changes link back to feedback or evidence?
- Are non-actions explained where appropriate?

11.3 Context preservation

- Is reputation scoped to zones and contribution domains?
- Does portable reputation carry rubrics, provenance, and limits?
- Are degraded signals clearly labeled?

11.4 Contestability and repair

- Can participants dispute inaccurate reputation signals?
- Are there appeal timelines and responsible stewards?
- Do reputation states decay, renew, or repair over time?

11.5 Anti-gaming and safety

- Are feedback loops bounded against amplification spirals?
- Are coordinated attacks, sybil behavior, and reputation farming mitigated?
- Are reporting systems protected against weaponization?

11.6 Privacy and proportionality

- Is reputation based on necessary signals rather than totalizing surveillance?
- Are sensitive feedback records protected?
- Can participants understand which signals affect rights, rewards, or access?

11.7 Inclusion

- Can marginalized and less-resourced participants provide feedback meaningfully?
 - Are feedback channels multilingual, asynchronous, and accessible?
 - Are affected groups visible in system learning without being exposed to retaliation?
-

12. Implementation Patterns

12.1 Reputation domains instead of reputation scores

Use separate reputation domains for different capabilities rather than one universal score.

Example domains:

- trusted annotator
- reliable bridge builder
- safe moderator
- responsive maintainer
- accurate fact-checker
- inclusive facilitator
- responsible AI operator

12.2 Confidence-weighted signals

Signals should include uncertainty. A verified expert correction, a peer endorsement, an anonymous concern, and a bot-like mass vote should not carry the same weight.

12.3 Receipts everywhere

Feedback, moderation, role changes, rewards, and reputation updates should produce receipts that can be audited without exposing unnecessary private data.

12.4 Decay by default

Signals should age unless renewed by current evidence. Decay prevents permanent lock-in and reduces the power of old or stale interactions.

12.5 Role renewal cycles

Reputation-linked roles should require periodic review or renewal, especially for stewards, moderators, grant reviewers, and high-impact AI operators.

12.6 Affected-group weighting

Feedback from those directly affected by a policy or harm may require special visibility or routing, while still protecting against capture.

12.7 Deliberative escalation

High-impact reputation changes should move from automated detection to human or community review before affecting access, compensation, or public standing.

12.8 Reputation portability bundles

Portable reputation should export as bundles containing claims, provenance, rubrics, expiry, dispute status, and interpretation limits.

12.9 Community retrospectives

Communities should periodically review whether feedback loops are actually improving governance, safety, inclusion, and contribution quality.

13. Relationship to Other Desirable Properties

DP1 – Identity and Accountability

DP18 depends on accountable identity to prevent sybil abuse, retaliation, and reputation laundering. DP1 provides the identity substrate; DP18 governs the memory of behavior and contribution.

DP2 – Participant Agency and Empowerment

Participants must understand and contest reputation effects. Feedback systems must produce real agency, not symbolic participation.

DP3 – Adaptive Governance

DP18 supplies the learning loops that allow governance to adapt. DP3 defines how decisions change; DP18 defines how lived experience becomes actionable signal.

DP4 – Data Sovereignty and Privacy

Feedback and reputation must operate under data minimization, purpose limitation, and consent. Reputation cannot justify surveillance.

DP7 – Interoperability

Reputation and feedback must preserve meaning across systems or degrade visibly.

DP8 – Community-Defined Participation and Governance Zones

Reputation is zone-scoped by default. Communities define which signals matter, which roles reputation unlocks, and how repair works.

DP9 – Developer and Community Incentives

DP18 provides the recognition and evaluation substrate for fair incentives. DP9 governs how value flows from contribution.

DP12 and DP13 – AI Governance and Containment

AI agents need scoped reputations, feedback-triggered containment, and human-ratified adaptation.

DP14 – Trust and Transparency

Feedback loops and reputation logic must be legible enough for participants to understand how trust is being shaped.

DP20 – Ownership and Stewardship

Communities should be able to own and govern their feedback data, reputation schemas, and contribution histories.

14. Non-Goals and Explicit Boundaries

DP18 is frequently misread as an argument for more measurement. It is not. The following are explicit non-goals, stated so that implementers cannot claim DP18 for systems it was written to prevent.

DP18 does not create a universal reputation score. There is no single number, tier, or trust level that follows a person across the meta-layer. Reputation is domain-scoped and zone-scoped by construction, and any system that collapses domains into one general standing has left DP18, not implemented it.

DP18 does not authorize social credit. Reputation may describe contribution within a bounded context. It may not be used to rank persons generally, to condition access to essential services, or to produce behavioral compliance through generalized scoring.

DP18 does not justify surveillance. Feedback and reputation are not a warrant for continuous telemetry. Signals must be necessary for a declared function, minimized, and governed under DP4. "We needed the data to compute trust" is not a defense.

DP18 does not make feedback binding. A signal is a legitimate input to a decision, not a decision. Communities remain free to decline requests, provided the decision and its reasoning are visible. DP18 requires loop closure, not majority rule by complaint volume.

DP18 does not replace governance. Feedback informs governance under DP3 and DP8; it does not substitute for deliberation, rights, or due process. Automated adaptation driven by signals must remain bounded and human-ratifiable.

DP18 does not prescribe one interface, schema, or algorithm. Zones may implement feedback objects, rubrics, and weighting differently. DP18 constrains the properties those implementations must preserve: typing, provenance, contestability, decay, and receipts.

DP18 does not promise neutrality. Any feedback system encodes choices about what counts. DP18 requires those choices to be published, versioned, and contestable rather than presented as objective measurement.

DP18 does not guarantee that harm will be caught. It requires that reporting exist, be safe, and be answerable. Absence of reports is not evidence of absence of harm, and no implementation may treat a quiet dashboard as a health claim.

DP18 does not permit permanent records as punishment. Accountability without a path to repair is exile. Retention that cannot decay, expire, or be contextualized is outside DP18 regardless of accuracy.

DP18 does not extend to identity verification. Establishing who someone is belongs to DP1. DP18 governs the memory of what was contributed and how it was experienced.

DP18 does not require that all reputation be portable. Portability is bounded by rubric fidelity. Where meaning cannot survive transfer, DP18 prefers visible degradation over silent translation (DP7).

15. Minimum DP18 Alignment (Non-Normative)

A system claiming DP18 alignment SHOULD demonstrate:

- structured feedback objects
- visible feedback routing and response commitments
- reputation scoped by zone and contribution domain
- published reputation logic where it affects access, rewards, or visibility
- decay, renewal, and repair mechanisms
- dispute and appeal pathways
- protection against sybil attacks, brigading, and reputation farming
- privacy-preserving data practices
- AI feedback processing with disclosure and human oversight
- adaptation receipts linking feedback to changes
- safeguards against runaway amplification loops
- explicit interoperability semantics for reputation portability

The following statements describe what each of those conditions means in practice. They are non-normative and are offered as an assessment aid rather than as a conformance test.

15.1 Signal integrity

Feedback is captured as typed, scoped objects with provenance and confidence metadata. Evidence, opinion, endorsement, complaint, and appeal are distinguishable. Low-confidence signals cannot trigger high-impact outcomes without review.

15.2 Loop closure

Every accepted signal has a state a participant can observe: received, routed, under review, acted upon, declined with reasoning, or deferred. Changes to rules or systems reference the signals that informed them.

15.3 Contextual memory

Reputation exists as named domains bound to zones and rubrics. No general-purpose score is exposed. Cross-zone use carries rubric, issuer, and limits, or is labeled as degraded.

15.4 Contestability and repair

Participants can inspect, annotate, and challenge attestations that describe them. Challenges have owners and timelines. Standing decays, renews, and can be repaired under published criteria.

15.5 Proportionate data practice

Collection is minimized and purpose-bound. Public recognition and private telemetry are separated. Reporters have protection appropriate to their risk, including pseudonymous pathways where safe.

15.6 Bounded automation

AI participation in clustering, summarization, or weighting is disclosed, sampled for accuracy, and ratified by humans before it affects rights, compensation, or access.

15.7 Anti-gaming posture

Amplification loops are damped. Coordinated inauthentic signaling, reputation farming, and weaponized reporting have named mitigations and observable incident handling.

15.8 Governance of the loop itself

The rules governing feedback and reputation are published, versioned, and amendable through the zone's governance process, with a record of who changed what and why.

16. Open Questions and Future Work

1. What minimum schema should define a feedback object across the meta-layer?
2. What minimum schema should define a reputation object?

3. Which reputation dimensions should be standardized, and which should remain zone-defined?
 4. How should reputation decay be represented across systems?
 5. What privacy-preserving methods can support reputation without centralized surveillance?
 6. How should communities distinguish endorsement, expertise, contribution, care, and popularity?
 7. What rights should participants have when reputation affects access, compensation, or visibility?
 8. How should non-human agents receive, carry, and contest reputation?
 9. What signals are safe to make portable across zones?
 10. How can marginalized communities govern feedback without being overexposed or tokenized?
 11. What forms of feedback should trigger mandatory governance review?
 12. How can reputation-linked roles avoid ossification and capture?
 13. What are the rollback standards for failed reputation experiments?
 14. How should feedback systems disclose AI summarization, clustering, or weighting?
-

17. Path Toward ML-RFC

DP18 is currently an ML-Draft and serves as exploratory scaffolding for how feedback loops and reputation can function as accountable learning infrastructure in the meta-layer.

Advancement toward ML-RFC status SHOULD require:

- convergence on a minimal interoperable schema for feedback objects
- convergence on a minimal interoperable schema for reputation objects
- demonstrated implementations across multiple zones
- tested mechanisms for decay, repair, and dispute resolution
- validated protections against gaming, sybil attacks, and amplification loops
- evidence of privacy-preserving reputation systems in practice
- working models for AI-assisted feedback synthesis with human oversight
- at least one cross-system portability experiment with semantic preservation
- documented governance processes for evolving reputation logic

ML-RFC promotion SHOULD be contingent on:

- rough consensus across participating communities
- demonstrated real-world usage with measurable learning outcomes

- clear failure cases and mitigation strategies
- alignment with DP1, DP2, DP3, DP4, and DP8 in deployed systems

Early ML-RFC candidates may focus on:

- feedback object standards
- reputation object standards
- decay and repair protocols
- feedback receipt and adaptation receipt formats

DP18 is expected to evolve iteratively, with partial RFCs emerging for specific components rather than a single monolithic specification.

18. Closing Orientation

DP18 makes the meta-layer capable of learning.

It ensures that feedback is not extracted and ignored, that reputation is not reduced to popularity, and that communities can recognize contribution without building permanent social cages.

A DP18-aligned meta-layer remembers enough to be accountable, forgets enough to allow growth, and adapts visibly enough for participants to trust that their experience matters.

Feedback becomes civic signal.

Reputation becomes contextual memory.

Learning becomes shared infrastructure.

This is the difference between a system that listens and one that evolves.
